AFRL-RH-WP-TR-2016-0075

Evaluation of Physiologically – Based Artificial Neural Network Models to Detect Operator Workload in Remotely Piloted Aircraft Operations

Michael Hoepf, Jonathan Mead,
Christina Guenwald, Paul Middendorf
Oak Ridge Institute for Science and Education

Matt Middendorf
Middendorf Scientific Services

Chelsey Credlebaugh
Ball Aerospace and Technologies

Scott Galster
Air Force Research Laboratory

July 2016

Interim Report

STINFO COPY

AIR FORCE RESEARCH LABORATORY
711 HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2016-0075 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signed//
Kyle Traver
Work Unit Manager
Applied Neuroscience Branch

//signed//
Scott M. Galster
Chief, Applied Neuroscience Branch
Warfighter Interface Division

//signed//
William E. Russell
Warfighter Interface Division
Airman Systems Directorate
711Human Performance Wing

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
13-07-16	Interim Report	1 August 2015 – 8 July 2016

4. TITLE AND SUBTITLE

Evaluation of Physiologically – Based Artificial Neural Network Models to Detect Operator Workload in Remotely Piloted Aircraft Operations

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H0DC (53273027)

6. AUTHOR(S)
**Michael Hoepf, +Matt Middendorf, **Jonathan Mead, **Christina Gruenwald, ++Chelsey Credlebaugh, **Paul Middendorf and Scott Galster*

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
**ORISE 4692 Millennium Drive, Suite 101Belcamp, Maryland 21017 ,+Middendorf Scientific Services 227 East Main Street, Medway Ohio 45341, ++Ball Aerospace 2875 Presidental Drive, Fairborn Ohio 45324

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory*
711Human Performance Wing
Airman Systems Directorate
Warfighter Interface Division
Applied Neuroscience Branch
Applied Adaptive Aiding Section
Wright-Patterson Air Force Base, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHCP

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2016-0075

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A: Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
88 ABW Cleared 10/31/2016; 88ABW-2016-5522. Report contains color.

14. ABSTRACT

The current research focuses on preventing performance decrements associated with mental overload during remotely piloted aircraft (RPA) operations. This can be accomplished using physiological signals to sense moments of high cognitive workload and providing augmentation to reduce workload and improve performance. Two RPA operators were interviewed to identify factors that impact workload in RPA, surveillance and target tracking missions. Performance, subjective workload, cortical, cardiac, respiration, voice stress, and ocular data were collected. Several physiological measures were sensitive to changes in workload as evidenced by performance and subjective workload data. In addition, several real-time workload models were evaluated. Potential future applications of this research include closed loop systems that employ advanced augmentation strategies, such as adaptive automation. By identifying physiological measures well suited for monitoring workload in a realistic simulation, this research advances the literature toward real-time workload mitigation in RPA field operations.

15. SUBJECT TERMS
Use unique descriptive terms to make the document findable in DTIC. Check DTIC Thesaurus

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kyle Traver
Unclassified	Unclassified	Unclassified	SAR	51	19b. TELEPHONE NUMBER *(Include Area Code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

TABLE OF CONTENTS

LIST OF TABLES

Page

ACKNOWLEDGMENTS

The authors would like to thank Kevin Durkee and Noah DePriest for their technical support, Captain Andrea Paulson-Metzger for support with data collection, Air Force Research Laboratory (AFRL) System Control Interfaces Branch (RHCI) for simulator support, AFRL Battlespace Acoustics Branch (RHCB) for software support, and Chuck Goodyear for assistance in statistical analysis.

This research was supported in part by Ball Aerospace and by an appointment to the Student Research Participation Program at the U.S. Air Force Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and USAFRL. The views expressed in this report are solely those of the authors and do not necessarily reflect the views of the employers or granting organizations.

1.0 SUMMARY

The goal of the current line of research is the prevention of performance decrements associated with mental overload during remotely piloted aircraft (RPA) operations. This can be accomplished using physiological signals as inputs to models that sense moments of high cognitive workload. When high workload is detected, augmentation can be provided to reduce workload and improve performance. Performance, subjective workload, cortical, cardiac, respiration, voice stress, and ocular data were collected in surveillance and target tracking missions. Several physiological measures were sensitive to changes in task load as evidenced by performance and subjective workload data. The primary focus of the current study was the evaluation of several real-time workload assessment models. Potential future applications of this research include closed loop systems that employ advanced augmentation strategies, such as adaptive aiding. By identifying physiological measures well suited for monitoring workload in a realistic simulation, this research advances the literature toward real-time workload mitigation in RPA field operations.

2.0 INTRODUCTION

U.S. armed forces are increasingly using RPA to accomplish missions in hostile environments because of their standoff capability in areas that are difficult to access or otherwise considered too hazardous for manned aircraft or personnel on the ground (U.S. Department of Defense, 2011). It has been documented that the military intends to increase the number of RPA in service while simultaneously reducing the number of operators (Dixon, Wickens, & Chang, 2004). One proposal to accomplish this is to allow operators to control multiple aircraft simultaneously (Rose, Arnold, & Howse, 2013). However, piloting one aircraft remotely is a complex task, and operating additional aircraft could increase task demands sharply. This is potentially problematic because cognitive overload can negatively impact performance (Young & Stanton, 2002). One solution to offset the risk of increasing the operator-to-vehicle ratio, as well as conventional sources of operator overload, is to monitor operator workload in real-time and provide augmentation before performance decrements occur. Physiological measures, which have been shown to reflect changes in cognitive workload in various environments (e.g., Wilson & Russell, 2007), are well suited for this goal. The current research is directed toward workload monitoring using physiological measures.

The Sense-Assess-Augment (SAA) framework was developed by researchers Galster and Johnson (2013) to detect and mitigate cognitive overload. This framework can be applied to a wide range of task domains. In general, the framework serves to *sense* the operator's physiological measures, *assess* their cognitive state using models or context-sensitive assessment tools, and *augment* the operator's performance before performance decrements may occur. The SAA framework was applied in a series of recent experiments using an RPA task environment.

First, physiological features that were sensitive to changes in workload were identified. Next, models were evaluated that utilized these features to assess the cognitive state of RPA operators. The goal of the model evaluation was a correlation between model output and subjective workload rating of $r \geq 0.85$. When the validated model detects cognitive overload (high workload), augmentation can then be implemented to reduce high workload and improve performance.

Before physiological measures can be used to monitor workload in RPA field operations, additional research is needed using realistic task environments. This is because the usefulness of each category of physiological measures (cortical, cardiac, etc.) for assessing workload likely depends on the nature of the task being performed. Hankins and Wilson (1998), for instance, found that cortical measures were sensitive to workload during mental calculation, cardiac measures were related to workload during flight segments heavily dependent on instrument use, and ocular activity was associated with workload during visually demanding flight segments. In addition, there is a need to develop an accurate physiologically-based workload model that operates in real-time with a high level of resolution. Wilson and Russell (2007), for instance, used an artificial neural network (ANN) model to monitor workload in real-time for an RPA task. However, their model only had a discrete (low vs. high) workload estimate. The current research attempts to build on this line of research by evaluating several models (see Durkee, Geyer, Pappada, Ortiz, & Galster, 2013) that provide a workload estimate on a continuous scale (0-100) within the context of a realistic RPA task environment. A continuous, rather than discrete, output affords the opportunity to better define thresholds necessary for augmentation control. A continuous scale will also provide more resolution for future applications that may benefit from intermediate levels of augmentation.

To address the need for the development of a realistic task environment in which to evaluate physiological measures and workload assessment models, a series of experiments were designed using a high-fidelity RPA simulation. Two RPA subject matter experts (SMEs) were interviewed to identify factors that can impact workload in surveillance and target tracking tasks. As an initial validation effort, two smaller scale studies were conducted based on the information obtained from these interviews. The first study utilized a surveillance task in which three factors identified by the SMEs (degraded sensor feed, number of distractor entities, and irrelevant communications) were combined in a within subjects factorial experiment. Subjective workload was higher and performance was lower in conditions with degraded sensor feeds and higher number of distractors. Thus, the degraded sensor feed and distractor manipulations were identified as effective, and were incorporated into the current research. These manipulations will be described further in the method section. The irrelevant communications manipulation involved the addition of extraneous communications, which were expected to increase workload by distracting participants from the task. The presence of these irrelevant communications did not impact subjective workload or performance, and thus the manipulation was not included in

the current experiment. Because this was an exploratory study, a small number of participants were used, and thus the study was not published. However, it was a very useful pilot study because two of the three experimental manipulations were found to have a significant effect on subjective workload.

In a follow up investigation (Hoepf, Middendorf, Epling, & Galster, 2015), a tracking task was developed and three workload factors identified by SMEs were incorporated into a within subjects factorial experiment. The manipulations were route type (country vs. city), number of RPA (one vs. two), and haze (off vs. on). City routes and two RPA conditions resulted in higher workload and reduced performance, and thus these manipulations were incorporated into the current research and are further described in the method section. The haze manipulation involved a change in the weather conditions. Settings in a virtual reality scene generator were used to create a hazy / foggy condition that was expected to make it more difficult to complete the task due to reduced visibility than the sunny / clear condition. The haze manipulation, however, did not significantly impact subjective workload or performance, and was therefore not used in the current research.

In the current investigation, the factors identified to drive workload in the two exploratory studies were implemented in a larger scale study. The surveillance task was combined with the tracking task into a single trial with a brief pause between the two tasks. During this pause, subjective workload ratings were collected. The task environment was then used to evaluate several previously developed physiologically driven ANN models of workload (see Durkee et al., 2013). This experiment also provided the opportunity to incorporate and evaluate new physiological features not available in the exploratory studies. The physiological measures utilized in the current research generally fall into five categories including cortical, cardiac, respiration, voice, and ocular measures.

2.1 Physiological Measures

Cortical, cardiac, respiration, voice, and ocular measures are all potentially well suited for monitoring workload. The physiological data needed to compute these measures can be collected in real-time, in a non-invasive manner, with devices and electrodes that are easy to apply. It is, however, important to note that in order to derive useful metrics from physiological data, signal processing algorithms are needed. Unprocessed electrical data from the eyes, for instance, is not necessarily useful for monitoring workload. However, this data can be processed to extract features (e.g., blink rate and duration) that have demonstrated sensitivity to changes in workload.

Cortical Measures. There are numerous neuroimaging techniques available for studying the complex and dynamic behavior of the brain. Electroencephalography (EEG) is employed in the current study because it offers high temporal resolution, ease of use, portability, and is of

relatively low cost compared to other neuroimaging techniques (Zander & Kothe, 2011). EEG is the recording of electrical activity along the scalp, which measures voltage fluctuations resulting from ionic current flows within the neurons of the brain (Niedermeyer & da Silva, 2004). Typical methods to examine EEG data include: power spectral density or the averaged power, maximum / log power spectra, sub-band entropy, and autoregressive modeling (Zarjam, Epps, & Lovell, 2012). Researchers have demonstrated that EEG can be used in real-time to assess mental workload (e.g., Wilson & Russell, 2007), and that such methods are sufficiently stable to provide accurate assessment over the course of several days and weeks (Christensen, Estepp, Wilson, & Russell, 2012). Research has shown that alpha activity is an idling rhythm of humans at rest, which becomes desynchronized during cognitive processes (e.g., higher workload; Pfurtscheller & Lopes da Silva, 1999). Thus alpha power should decrease in high workload conditions (Wilson, 2001). Conversely, theta and delta power have been found to increase under high workload conditions (Hankins & Wilson, 1998; Wilson, 2001).

Cardiac Measures. Electrocardiography (ECG) can be used to obtain cardiac measures, such as heart rate (HR) and heart rate variability (HRV), in most task environments via the application of a few electrodes over the heart (Wilson, 1992). In both laboratory and field settings, researchers typically observe HR increases and HRV decreases in high workload situations (e.g., Jorna, 1992; Mulder, 1992; Porges & Byrne, 1992; Roscoe, 1992). There is some debate about which measure is superior. Roscoe (1992) suggested that HRV may indicate changes in mental workload in the absence of any change in overall HR. Grossman (1992), however, indicated that it is not clear if HRV provides any more information than simple HR.

Respiration Measures. An inexpensive, simple, and non-invasive method to collect respiration data involves the operator wearing an elastic band around the rib cage which measures expansion associated with breathing. Respiration has been clinically associated with the autonomic nervous system (ANS), and as such is affected by the body's stress response (Suess, Alexander, Smith, Sweeney, & Marion, 1980). As stress and workload increase, so do the metabolic demands within the body. Breathing rate is the most frequently utilized respiratory variable in psychophysiological research today, but it may not be the best indicator of stress. Cohen et al. (1975) suggested that one must break the respiratory waveform down into its various components and analyze them separately in order to visualize the precise effects of the stress response on respiratory physiology. Further, Veltman and Gaillard (1998) indicate that inspiratory flow should increase, while inspiration time, respiration amplitude, respiration cycle time, and respiration duty cycle time should decrease, during times of high mental workload.

Voice Measures. Of all of the physiological measures described thus far, voice data is perhaps the easiest to obtain. Operators typically utilize headsets equipped with microphones, so collecting voice data is simply a matter of utilizing the microphone. The more challenging aspect of voice stress analysis is the extraction of the specific features that are associated with

workload. Certain vocal properties allow for the non-invasive diagnosis of psychophysiological state in real-time (Collier, 1974). Fundamental frequency measures include mean pitch, pitch variance, maximum pitch, and pitch range, which have all been shown to increase with larger levels of cognitive workload (Brenner, Doherty, & Shipp, 1994; Lippold, 1971). Additional voice stress measures include speaking rate (syllables per second), average syllable length, average pause length, and percent pause (i.e., the percentage of time spent pausing in an utterance). Speaking rate and percent pause have been shown to increase under higher cognitive loads, while average syllable and pause length decrease (Brenner et al., 1994).

Ocular Measures. Data from the eyes can be obtained via both on body methods, such as electrooculography (EOG), and off body methods, such as camera-based eye-tracking systems. Several ocular measures have demonstrated sensitivity to workload. Blink rate and blink duration, for instance, typically decrease with an increase in cognitive load (Fogarty & Stern, 1989). Furthermore, an increase in pupil diameter often occurs during an increase in mental demand (Beatty, 1982). It should be noted, however, that pupil dilation can also change due to the illumination condition of the visual field. In fact, background brightness can result in greater variation of pupil diameter than task difficulty (e.g., Pomplun & Sunkara, 2003).

Summary. In the current research, raw physiological data were collected, algorithms were used to extract features, and the features were used as inputs to ANN models (see Durkee et al., 2013). The models were evaluated and validated using a realistic high-fidelity RPA task environment. More specifically, the primary goal of the current research was to determine if the model outputs correlate with a validated measure of subjective workload.

Another objective of the current research was to identify new physiological features that show sensitivity to workload. Not all physiological features in the current research were used as model inputs, but all features were evaluated. The respiration and voice measures in this study, for instance, were not input into the models, but were nonetheless analyzed in relation to the workload manipulations. The following experiment was designed to meet these research objectives.

3.0 METHODS

3.1 Participants

Twelve individuals recruited locally (Midwest region) were studied. Eight participants were male and four were female. Age ranged from 18-46, with a mean of 25.7. They were screened for motor, perceptual, cognitive, heart, and neurological conditions, as well as hearing impairments. Participants did not take any neurological medications or medications that caused drowsiness. The participants stated they were comfortable operating a computer, reading small characters on

a computer monitor, hearing and comprehending verbal commands presented through headphones, and learning complex, computer based tasks. They were fluent in English and had normal or corrected-to-normal eyesight with no color blindness, and provided written informed consent in accordance with human research ethics guidelines prior to the start of the experiment. These participants were non-pilots, lacked operational experience, and were paid for their participation.

In addition to the general sample, two pilots participated in the study. Data from the pilots were examined, though not included in the general analysis due to their operational experience. Both pilots were male. One was 38, and the other was 35 years old. One had 7 years of RPA flight experience, and the other had 3 years. The pilots volunteered to participate in the study and were not paid for their time. The same screening applied to the general sample also applied to the pilots.

All study procedures were reviewed and approved by the Air Force Research Laboratory Institutional Review Board.

3.2 Task

Task overview. Each trial consisted of two separate primary tasks, both of which coincided with a secondary communications task (see Appendix A for screenshots of the tasks). Both primary tasks were implemented on a RPA simulator called "Vigilant Spirit." This software was produced by the Air Force Research Laboratory (AFRL) System Control Interfaces Branch (RHCI). The secondary task was created using the Multi-Modal Communications (MMC) tool. This software was produced by the AFRL Battlespace Acoustics Branch (RHCB). Trials were presented to the participants as a simulated mission that started with a surveillance task and was followed by a tracking task. From a research perspective, this structure consists of two separate 2 x 2 factorial experiments. There were 16 scenarios that each participant experienced once over the course of four days of data collection. Conditions were counterbalanced within task type (surveillance and tracking), although the tracking task always followed the surveillance task.

Each trial would begin with one minute of setup time. The surveillance task followed this setup phase, taking place between 60 - 315 seconds into each trial. At 330 seconds, an audio message was presented over the headset indicating that it was time for participants to complete a subjective workload questionnaire (described later) for the surveillance task. At 490 seconds into the trial, another message played indicating that it was time for the participant to prepare for the tracking task. Participants were then guided through a target acquisition phase, followed by the tracking task which took place between 600 - 810 seconds. A message in the headset then indicated that the trial was over and that it was time for the participant to complete a subjective workload assessment of the tracking task.

Surveillance task. The surveillance task required participants to search a market area to find high value targets (HVTs). The HVTs were men carrying M82 sniper rifles. There were four HVTs per trial and they appeared at one minute intervals. The HVTs entered the scenario by walking out from under a tent and walked around the market area for roughly 57 seconds. They exited the scenario by walking under a different tent, at which point a new HVT would enter the scene. The four HVTs never overlapped.

The first experimental manipulation was the number of distractors (other entities) walking around the market area. There were three types of distractors: women wearing long dresses, men carrying pistols, and men carrying shovels. We expected that the male distractors would present a larger challenge than the female distractors because of their visual similarity to the HVTs (same entity model, but carrying something different). We also expected that the men carrying the shovels would be especially distracting because the shovel was visually similar to the sniper rifle, as was held in a position similar to the sniper rifle. The easy conditions contained 16 distractors (8 women, 7 men with pistols, and 1 man with a shovel), and the hard conditions contained 48 distractors (24 women, 20 men with pistols, and 4 men with shovels).

The second experimental manipulation was the presence or absence of fuzz (degraded sensor image). Under the easy condition the fuzz was not turned on, but under the hard condition the fuzz was on, making it more difficult to identify HVTs. The image degradation was designed to reflect equipment failure that can occur in the real world and was similar to the "snow" seen on television sets when the signal is not clear.

Tracking task. For this task, participants were required to track HVTs traveling by motorcycle. Participants were instructed to keep the RPA sensor positioned over the HVTs, which they accomplished by clicking in the sensor feed with the mouse, causing the sensor feed to center on where they had clicked. A feature in the RPA simulator called sensor slaved tracking would then automatically update the aircraft position to fly a loiter circle around the center of the sensor feed, thereby eliminating the need for the participant to manually navigate the aircraft.

The first experimental manipulation was the number of HVTs. In easy conditions participants tracked one HVT, and in hard conditions they tracked two. Tracking two HVTs was expected to be more difficult because it required participants to constantly shift their attention between two video feeds.

The second experimental manipulation was the route the HVT(s) would travel. In easy conditions the HVTs would ride along a straight, open, country road. In hard conditions HVTs would travel through a city, taking many turns and frequently becoming obstructed by buildings.

Communication task. A secondary task was presented concurrently with both of the primary tasks. The task consisted of answering four cognitively challenging mental math questions. Questions were presented verbally over a headset and transcriptions were displayed. Participants were instructed to press and hold the spacebar while they responded verbally. The questions were operationally relevant. An example question is "How long would it take you to reach a destination 100 nautical miles away with a headwind of 15 knots?" Questions were evenly distributed throughout each trial.

3.3 Apparatus and Measures

Performance. Performance was assessed using a composite scoring algorithm, which was based on performance in both the primary and secondary tasks. The maximum possible score was 1,000 points (800 for the primary task and 200 for the secondary task). This basic approach is used for both the surveillance and tracking tasks.

To obtain points in the surveillance task, participants were required to locate, identify, and track the HVTs. Participants pressedthe F-key when they thought they had found the HVT. If correctly identified, points would begin accumulating for each second the HVT was tracked until he walked under a tent. Incorrectly identifying a distractor as the HVT (false positive) would result in a five point penalty. Participants were required to keep the HVT visible in the video feed while tracking to accrue points. Additionally, participants would receive the maximum number of points per second for keeping the video feed at one of the two highest zoom levels. Using lower zoom levels would result in half as many points being awarded.

Performance in the tracking task was divided into two components. First, participants were required to keep the HVTs, traveling by motorcycle, in the video feed(s). Maximum points were accumulated when using the highest two levels of zoom, whereas points accrued at half that rate at lower levels of zoom. At most, 600 points per trial could be attained from keeping the HVT(s) in the video feed. Second, participants were instructed to keep the HVT(s) centered in the video feeds. At most, 200 points per trial could be attained from keeping the HVT(s) centered.

For the secondary task, there were four questions per trial, each worth a maximum of 50 points. In order to obtain all points, participants had to respond correctly within 20 seconds. After 20 seconds, the participants would lose 5 points per second for the next 10 seconds. After 30 seconds, no points were given. Answering incorrectly resulted in a 5 point penalty. The four questions were evenly spaced within each task (surveillance & tracking). There were four groups of questions (speed addition, speed subtraction, distance, and altitude), with eight unique questions pergroup. Each trial consisted of a unique combination of one question from each group. Therefore, each question was presented twice over the course of data collection. Question difficulty was balanced across conditions.

Competition. During early in-house testing, one of the test subjects stated that they had simply given up in the hardest condition, and rated their subjective workload accordingly. This type of disengagement would have deleterious effects on the current study if a process was not implemented to discourage it. Therefore, a competition-based plan using performance scores was implemented. Prior research has shown that competition is an effective way to increase subject motivation in long-term experiments (Fiorita, Middendorf, & McMillan, 1992). Top scores based on session averages were posted on a whiteboard. Session averages were used to discourage participants from giving up on any one trial. The top scores were posted on the whiteboard using identification numbers to maintain anonymity.

To provide participants a better chance to post high scores and thus remain engaged, the whiteboard was periodically erased throughout the study. The decision to erase the board was motivated by the performance of the first participant, who posted the highest scores in the study. Thus, the goal of erasing the board was to prevent other participants from withdrawing from the competition because they did not believe that they would be able to get on the board, thereby negating the incentive to do their best.

Subjective workload. Subjective workload was collected using a modification of the NASA-Task Load Index (TLX), a multidimensional measure that assesses perceived workload (Hart & Staveland, 1988). Workload was determined by averaging across the six sub-scales (mental demand, physical demand, temporal demand, performance, effort, and frustration). Nygren (1991) found the average to be psychometrically equivalent to the weighted sub-scale averaging suggested by the TLX authors. Empirically, the weighted averages have not been found to be superior to the simple average of the sub-scales (Christ et al., 1993; Hendy, Hamilton, & Landry, 1993).

Physiological data acquisition and processing. The physiological data were collected using four hardware devices, including two Cleveland Medical Devices BioRadio 150s, a SmartEye Pro 5.9 eye-tracking system, and a Zephyr Bioharness 3. Electrical signals (EEG, EOG, and ECG) were sampled using the two BioRadios. All signals connected to the BioRadio 150 were subjected to a first order analog band pass filter with an input bandwidth of 0.5 - 250 Hz. The sampled data were transmitted wirelessly to a computer for processing and recording.

EEG data, sampled at 480 Hz, were acquired using electrodes placed directly on the scalp and secured in place with an Electro-Cap manufactured by Electro-Cap International, Inc. The EEG data were measured at seven sites on the scalp in accordance with the international 10 / 20 system (Jasper, 1958). The seven sites were F7, F8, T3, T4, Fz, Pz, and O2. The right and left mastoids were used as the reference and ground for the EEG signals. All initial electrode impedances were measured to be at or below 5 kΩ. The frequency bands (i.e., pass bands) used in the EEG signal processing were delta (1-3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-30

Hz), gamma 1 (31-40 Hz), gamma 2 (41-57 Hz) and gamma 3 (63-100 Hz). A four second time domain window was used to process the raw EEG data. The raw data in the four second window was filtered using a 4th order Butterworth band pass filter with break frequencies describe above. A Hanning window was applied to the filtered data and power spectral analysis was performed. The resulting power in the pass band was then averaged. These steps were repeated for each frequency band and electrode site. The four second time domain windows had a 75% overlap, thus yielding one measure of average power every second. This signal processing approach produced 49 EEG measures per second (7 sites with 7 bands per site). Artifact mediation was accomplished using the artifact separation technique (see Credlebaugh, Middendorf, Hoepf, & Galster, 2015).

Vertical EOG (VEOG) data were acquired using two electrodes placed above and below the left eye. Horizontal EOG (HEOG) data were acquired using two electrodes placed to the left and right of the eyes. All EOG data were sampled at 480 Hz, and the left mastoid was used as the ground. The initial electrode impedances for the EOG were measured to be at or below 20 kΩ. Blink rate and duration were extracted from the VEOG data using a blink detection algorithm (see Epling et al., 2015). Both the VEOG and HEOG were used to detect saccades (see Middendorf et al., 2015).

ECG data, sampled at 960 Hz, were acquired using two electrodes placed on the sternum and xiphoid process. The left mastoid was used as the ground. The initial electrode impedances for the ECG were measured to be at or below 20 kΩ. Interbeat intervals (IBIs) were calculated from the ECG data. The IBIs were used to calculate heart rate and heart rate variability.

Pupil diameter data were sampled at a frequency of 60 Hz using the SmartEye Pro 5.9 eye-tracking system. The system was comprised of six cameras mounted with infrared light sources, a computer used for the processing of image data, and the SmartEye software. It is important to note that there was a data quality variable associated with pupil diameter data. The variable ranged between 0 and 1, with 1 indicating highly reliable data, and 0 indicating that the data cannot be used. For example, if the participant blinked or left the field of view, the quality variable would be zero. The specific information used to formulate this variable was not available as the algorithm is proprietary. However, a conversation with a representative from the SmartEye organization indicated that a value of .65 would be an acceptable cutoff value in the determination of the usability of the data. Thus, only pupil diameter data associated with a quality variable value of .65 or higher were included in analysis.

The Zephyr Bioharness 3 was used to acquire respiration data. This device consists of two components, including an elastic strap and a data acquisition "puck." The strap is worn around the torso at the sternum level and the puck, which snaps into the strap, contains a microprocessor

for data acquisition and wireless transmission to a computer via Bluetooth. Respiration is captured as a breathing waveform signal using a capacitive pressure sensor, sampled at 18 Hz.

The breathing waveform signal is processed by an algorithm to produce six features as described by Veltman and Gaillard (1996). The six features are respiration rate, inspiration flow, inspiration time, expiration time, total cycle time, and duty cycle time, The algorithm is robust to occasional dropouts in the Bluetooth signal and artifacts caused by body movement.

Workload models. Four artificial neural network models were created using NeuroSolutions software. The goal of the models was to produce a real-time measure / estimate of cognitive workload based on physiological features. The models outputs ranged from 0-100, with 0 being the lowest and 100 being the highest workload estimate. The models were initially trained using data from a previous investigation (see Durkee, Pappada, Ortiz, Feeney, & Galster, 2015). All models utilized 42 EEG inputs (including all of the 49 EEG measures listed above with the exception of all of the bands at the T4 site), pupil diameter, heart rate, and inter-beat interval. The T4 site could not be included as a model input because it was not included in the training dataset.

Model 1 operated with static weights from a recent model training study (Durkee et al., 2015). The complete dataset consisted of 1,875 minutes of physiological data. The data were collected from 25 participants in a prior study as they completed 15 trials that lasted 5 minutes each. In addition, the model was trained via ANN methods in which the inputs were collectively trained to an estimate of ground truth. The ground truth estimate was derived from the continuous time series using the ratio of Fz theta / Pz alpha for signal injection. Gevins (1998) found that frontal theta power increases with task load while parietal alpha power decreases. Furthermore, this EEG ratio was anchored to participants' NASA-TLX responses for each corresponding trial, meaning the average continuous truth estimate for each trial approximately matched the NASA-TLX response.

A brief description of the signal injection process using the ratio is provided here. Essentially, a validated metric (the TLX in this case) is collected and used as an anchor. Since the TLX is only collected at the end of the trial, another validated workload metric that can be used continuously is used to retrospectively inject noise (e.g., the Fz / Pz EEG ratio). The output of this signal injection step is a second-by-second estimate of "actual workload," which if averaged over the course of a given trial, would roughly equal the TLX response for the trial. Noise is injected into the TLX under the assumption that workload does not remain perfectly constant throughout a task, and the TLX is merely an average workload estimate by a person over the task. Using the TLX alone would circumvent efforts to estimate specific workload fluctuations, including the peaks and valleys. The signal injection step attempts to objectively define these fluctuations. It does this by tracking an objective measure that has been empirically shown to have a relationship

with workload. Although this does not give a perfect estimate of workload, the hypothesis is that it should resemble actual workload better than assuming a flat-line TLX across the entire trial, and training the model to find the best fitting weights to match that. All four models used some sort of signal injection, although two of the models use Fz theta only for signal injection instead of the ratio. The other important distinction with respect to the models is whether static weights were used versus on-line training.

Model 2 was identical to Model 1, but it trained on-line. On-line training in this experiment essentially means that the measurement system was actively and autonomously adapting / changing in an effort to better "learn" what it is trying to measure (i.e., reduce model error, improve classifier accuracy). More specifically, the model weights attempted to re-train each time a new TLX response was received while the data collection system was in a stopped state after a trial had ended. There were a number of reasons that the model weights would not re-train. For instance, if model training was currently underway from a previous trial, this process would not be interrupted. Another reason model weights did not always update would be to retain existing model weights that provided an accurate classification relative to NASA-TLX responses. This was implemented through a system check comparing the mean difference between the NASA-TLX response and the output from the previous trial. Specifically, if the mean difference was less than five the model training process would be bypassed in favor of the existing weights.

For all on-line model training attempts, the TLX input served as a trigger for the system to append the newly received data to the overall model training set and to initiate a new model training process. One important caveat is that updated participant weights from on-line training were not saved from session to session. Thus, the outputs of Model 1 and 2, for instance, would be identical for the first trial of each session, but would diverge for the remaining three trials for each session. It should be noted that the models that trained on-line were based on 84% smaller datasets than the static weight models. Although a reduced dataset can impact model accuracy, this was necessary to accommodate reasonable training times, in the range of four to six minutes on average. In addition, the smaller datasets were intended to increase the sensitivity of the new model weights to each subjects' patterns in their respective physiological signals, thus, more rapidly creating model weights that are personalized to a specific subject. Although the additional data used in the on-line training process was relatively small compared to the initial training dataset, the goal was to obtain a glimpse of how on-line training may improve model accuracy.

Models 3 and 4 were different from Models 1 and 2 in that they utilized Fz theta only for signal injection instead of the ratio. Prior research (e.g., Onton, Delorme, & Makeig, 2005) has shown an increase in frontal midline theta activity with increased working memory demands. It was important to investigate multiple types of model configuations in order to determine if one

method was superior. In addition, Model 3 utilized static weights whereas Model 4 trained on-line. Examining these various model development strategies allowed for a direct comparison of model accuracy as a function of using different data features as well as the impact of individualized models.

3.4 Procedure

Participants were brought into the laboratory for two days of training and four days of data collection. On the first training day, participants viewed a PowerPoint presentation containing a description of the task and measures, and then completed part-task training for the primary and secondary tasks. The first day of training concluded with four practice trials. The second day of training consisted of a minimum of four additional practice trials, with the possibility of running up to an additional two practice trials. There were two reasons why participants would sometimes complete additional practice trials. First, extra practice trials were provided if a participant requested extra practice. Second, the research team mandated additional practice trials if a participant struggled to consistently meet a minimum performance threshold. This minimum performance threshold was defined as the ability to consistently obtain points in both the primarily and secondary task, while demonstrating a thorough understanding of the composite scoring algorithm. On data collections days, participants were equipped with the physiological measurement devices and then completed four experimental trials per day, for a total of sixteen trials. A debriefing was conducted at the end of the last day.

4.0 RESULTS

There were two primary areas of analysis from the experiment, the ANOVA and model results. First, descriptive and ANOVA results are presented, followed by the model results for the surveillance task. The same structure will follow for the tracking task. Formal statistical analyses were only conducted using data from the general sample, as the two pilots did not constitute a sample size sufficient for statistical analyses.

4.1 Surveillance Descriptive and ANOVA Results

The means from the surveillance task are presented in Table 1 (distractor manipulation) and Table 3 (fuzz manipulation). The two pilots showed similar physiological responses to the non-pilots. The means for the performance and subjective workload measures for the pilots are presented in Table 2 for the distractor manipulation and Table 4 for the fuzz manipulation.

Table 1. Means for the distractor manipulation for the surveillance task.

Variable	Low Distractors		High Distractors	
	Mean	**SE**	**Mean**	**SE**
Primary Score	449.833	17.826	310.082	23.860
Number of HVTs Located	3.281	0.120	2.385	0.141
Secondary Score	181.625	5.250	181.031	5.076
Total Score	631.458	20.787	491.113	26.085
Subjective Workload (TLX)	37.404	4.134	42.934	4.025
Heart Rate (Beats / Minute)	72.370	3.386	72.068	3.680
Heart Rate Variability (Hz)	0.028	0.124	0.028	0.117
Inspiration Flow	2.245	0.145	2.284	0.153
Inspiration Time (s)	0.221	0.026	0.225	0.029
Respiration Amplitude	519.369	1.269	519.747	1.411
Respiration Cycle Time (s)	3.116	0.104	3.137	0.107
Respiration Duty Cycle Time (s)	0.413	0.006	0.412	0.004
Mean Pitch (Hz)	144.852	14.436	145.092	14.090
Pitch Variance (Hz)	99.285	0.402	112.830	0.392
Maximum Pitch (Hz)	173.974	19.214	174.764	18.509
Pitch Range (Hz)	41.112	0.206	42.976	0.195
Speaking Rate (syllables / second)	4.412	0.135	4.404	0.146
Average Syllable Length (s)	0.156	0.007	0.156	0.006
Average Pause Length (s)	0.133	0.005	0.131	0.005
Percent Pause (%)	38.359	1.596	38.185	1.572
Blinks Rate (Blinks / Minute)	9.270	0.201	8.808	0.197
Blink Duration (s)	0.107	0.003	0.107	0.003
Pupil Diameter (mm)	4.274	0.219	4.270	0.218

Table 2. Means for the distractor manipulation for the surveillance task among the pilots.

Variable	Low Distractors	High Distractors
Primary Score	368.121	288.916
Number of HVTs Located	2.625	2.188
Secondary Score	167.188	167.188
Total Score	535.308	459.229
Subjective Workload (TLX)	18.334	16.667

Table 3. Means for the fuzz manipulation for the surveillance task.

Variable	No Fuzz		Fuzz	
	Mean	**SE**	**Mean**	**SE**
Primary Score	387.077	23.028	372.839	17.269
Number of HVTs Located	2.896	0.126	2.771	0.118
Secondary Score	183.344	4.633	179.312	5.832
Total Score	570.421	25.532	552.151	19.182
Subjective Workload (TLX)	40.468	4.317	39.870	3.931
Heart Rate (Beats / Minute)	71.931	3.574	72.507	3.514
Heart Rate Variability (Hz)	0.028	0.107	0.028	0.135
Inspiration Flow	2.340	0.145	2.189	0.157
Inspiration Time (s)	0.217	0.029	0.229	0.027
Respiration Amplitude	520.045	1.378	519.071	1.341
Respiration Cycle Time (s)	3.094	0.111	3.159	0.106
Respiration Duty Cycle Time (s)	0.414	0.005	0.411	0.006
Mean Pitch (Hz)	145.542	14.540	144.402	13.990
Pitch Variance (Hz)	101.351	0.408	110.531	0.386
Maximum Pitch (Hz)	174.178	19.148	174.560	18.588
Pitch Range (Hz)	41.009	0.209	43.083	0.191
Speaking Rate (syllables / second)	4.406	0.140	4.409	0.141
Average Syllable Length (s)	0.156	0.006	0.156	0.007
Average Pause Length (s)	0.129	0.005	0.135	0.006
Percent Pause (%)	38.249	1.686	38.295	1.442
Blinks Rate (Blinks / Minute)	9.017	0.201	9.055	0.205
Blink Duration (s)	0.107	0.003	0.107	0.003
Pupil Diameter (mm)	4.284	0.219	4.260	0.218

Table 4. Means for the fuzz manipulation for the surveillance task among the pilots.

Variable	No Fuzz	Fuzz
Primary Score	316.639	340.398
Number of HVTs Located	2.438	2.375
Secondary Score	181.563	152.813
Total Score	498.201	496.336
Subjective Workload (TLX)	15.313	19.688

The performance, subjective workload, and physiological data were statistically evaluated using two-way repeated-measures ANOVAs for the surveillance task (see Table 5). Note that positively skewed dependent variables (e.g., blink rate) were log transformed prior to analyses. Figures for significant results can be found in Appendix B and EEG results are presented in Appendix D.

Table 5. ANOVA results for the surveillance task.

Variable	Distractors	Fuzz
Primary Score	$F(1,11) = 37.62$, $p = 0.0001$*	$F(1,11) = 0.50$, $p = 0.4924$
Number of HVTs Located	$F(1,11) = 43.00$, $p = 0.0001$*	$F(1,11) = 1.69$, $p = 0.2199$
Secondary Score	$F(1,11) = 0.03$, $p = 0.8662$	$F(1,11) = 1.01$, $p = 0.3823$
Total Score	$F(1,11) = 33.47$, $p = 0.0001$*	$F(1,11) = 0.83$, $p = 0.4049$
Subjective Workload (TLX)	$F(1,11) = 35.03$, $p = 0.0001$*	$F(1,11) = 0.14$, $p = 0.7108$
Heart Rate (Beats / Minute)	$F(1,11) = 0.30$, $p = 0.5932$	$F(1,11) = 0.63$, $p = 0.4447$
Heart Rate Variability (Hz)	$F(1,11) = 0.92$, $p = 0.3573$	$F(1,11) = 0.11$, $p = 0.7461$
Inspiration Flow	$F(1,11) = 1.78$, $p = 0.2185$	$F(1,11) = 6.01$, $p = 0.0399$*
Inspiration Time (s)	$F(1,11) = 0.56$, $p = 0.4746$	$F(1,11) = 1.00$, $p = 0.3456$
Respiration Amplitude	$F(1,11) = 3.57$, $p = 0.0954$	$F(1,11) = 4.12$, $p = 0.0770$
Respiration Cycle Time (s)	$F(1,11) = 1.06$, $p = 0.3329$	$F(1,11) = 1.72$, $p = 0.2262$
Respiration Duty Cycle Time (s)	$F(1,11) = 0.12$, $p = 0.7410$	$F(1,11) = 0.42$, $p = 0.5343$
Mean Pitch (Hz)	$F(1,11) = 0.09$, $p = 0.7666$	$F(1,11) = 1.30$, $p = 0.2781$
Pitch Variance (Hz)	$F(1,11) = 4.24$, $p = 0.0640$	$F(1,11) = 1.72$, $p = 0.2160$
Maximum Pitch (Hz)	$F(1,11) = 0.65$, $p = 0.4386$	$F(1,11) = 0.08$, $p = 0.7804$
Pitch Range (Hz)	$F(1,11) = 1.62$, $p = 0.2293$	$F(1,11) = 1.80$, $p = 0.2063$
Speaking Rate (syllables / second)	$F(1,11) = 0.02$, $p = 0.8801$	$F(1,11) = 0.00$, $p = 0.9581$
Average Syllable Length (s)	$F(1,11) = 0.01$, $p = 0.9127$	$F(1,11) = 0.06$, $p = 0.8063$
Average Pause Length (s)	$F(1,11) = 0.99$, $p = 0.3409$	$F(1,11) = 2.94$, $p = 0.1145$
Percent Pause (%)	$F(1,11) = 0.06$, $p = 0.8093$	$F(1,11) = 0.01$, $p = 0.9333$
Blinks Rate (Blinks / Minute)	$F(1,11) = 1.80$, $p = 0.2067$	$F(1,11) = 0.00$, $p = 0.9636$
Blink Duration (s)	$F(1,11) = 0.45$, $p = 0.5160$	$F(1,11) = 0.00$, $p = 0.9674$
Pupil Diameter (mm)	$F(1,11) = 0.06$, $p = 0.8040$	$F(1,11) = 0.41$, $p = 0.5366$

Note: * = $p < .05$; There were no significant interactions in the surveillance task.

Cortical measures. The EEG measures (power at each site and frequency band) were analyzed for each manipulation, but for conciseness only the significant ($p < .05$) results are reported and the means, standard errors, and F values are not included. Additionally, due to the fact there is little, or no, literature (Borghini, Astolfi, Vecchiato, Mattia, & Babiloni, 2014) reporting usefulness of the upper frequency bands (beta and gamma), these bands are not reported here.

For the fuzz manipulation, there was more power in the theta band at the F8 site in fuzz conditions than clear conditions. This effect was no longer present when blink and saccade artifacts were removed. A similar finding occurred for power in the delta band at site O2 (see Appendix D).

The findings for the number of distractors manipulation were much more interesting. Six of the seven sites in the delta band had significantly less power for a high number of distractors than low distractors. However, all six sites lost significance when blink artifacts were removed (see Appendix D). This is not surprising given that participants tend to blink less in high workload conditions than low conditions (Fogarty & Stern, 1989), and it is known that blinks easily contaminate the delta band (Picton et al., 2000). Although blink rate was not significant in this study ($p = 0.2067$), there were substantially less blinks in the high distractors conditions than the low distractors condition (4597 vs. 4859).

For the theta band there was significantly more power at sites T3 and Pz for high distractors than low. These results are in the expected direction. For the alpha band there was significantly more power at all seven sites for high distractors than low. This effect is not in the expected direction. It was suspected that the effect was caused by ocular artifacts, since these artifacts easily contaminate alpha power (Picton et al., 2000). When blink artifacts were removed, O2 lost significance. However, when blink and saccade artifacts were removed, three additional sites (mostly posterior) lost significance. The likely explanation for this is the change in the scanning pattern used for high and low distractors. For high distractors, the participants need to examine more entities to find the HVT, and the distance between the entities is smaller. This coupled with the fact that the saccade detection algorithm is threshold-based explains the effect. Specifically, the saccade detection algorithm is good at finding big saccades and less effective for small ones. Therefore, many of the smaller saccades associated with the high distractor condition do not get detected and the associated EEG measures do not get removed. The fact that only the anterior sites remain significant is not surprising given that as ocular artifacts travel across the scalp from anterior to posterior sites, the amplitude of the artifact decreases (Picton et al., 2000).

4.2 Surveillance Model Results

In order to evaluate the workload models, the relationships between the model outputs (averaged over the course of each trial) and Average TLX were investigated. Pearson partial correlations controlling for participant were calculated to determine if there was a relationship between Average TLX and model output. No correlations were significant (see Table 6), indicating that none of the models exhibited a relationship with subjective workload in the surveillance task.

Table 6. Pearson partial correlations between average trial output from ANN models and average subjective workload (TLX) in the surveillance task.

Model	R-value
1	-.03
2	.05
3	-.10
4	.22

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

4.4 Tracking Descriptive and ANOVA Results

The means from the tracking task are presented in Table 7 (target manipulation) and Table 9 (route manipulation). Means from the pilots for the performance measures and subjective workload are also presented (see Table 8 for the target manipulation and Table 10 for the route manipulation).

Table 7. Means for the target manipulation for the tracking task.

Variable	One Target		Two Targets	
	Mean	**SE**	**Mean**	**SE**
Center Score	149.022	1.037	128.941	3.295
Score Foot Print	594.392	2.722	553.862	14.373
Secondary Score	186.018	4.318	180.399	7.212
Total Score	930.050	5.811	863.656	19.907
Subjective Workload (TLX)	35.868	4.039	46.947	4.476
Heart Rate (Beats / Minute)	73.683	3.676	75.480	3.966
Heart Rate Variability (Hz)	0.027	0.121	0.026	0.125
Inspiration Flow	2.261	0.143	2.285	0.167
Inspiration Time (s)	0.236	0.024	0.219	0.026
Respiration Amplitude	519.696	1.209	519.634	1.389
Respiration Cycle Time (s)	3.170	0.103	3.065	0.100
Respiration Duty Cycle Time (s)	0.413	0.005	0.417	0.004
Mean Pitch (Hz)	144.532	14.094	145.543	14.137
Pitch Variance (Hz)	104.338	0.383	103.337	0.400
Maximum Pitch (Hz)	172.951	18.414	174.701	18.324
Pitch Range (Hz)	41.703	0.195	41.754	0.205
Speaking Rate (syllables / second)	4.301	0.119	4.341	0.151
Average Syllable Length (s)	0.158	0.006	0.160	0.007
Average Pause Length (s)	0.137	0.005	0.136	0.005
Percent Pause (%)	38.053	1.516	35.874	1.349
Blink Rate (Blinks / Minute)	8.589	0.250	6.663	0.246
Blink Duration (s)	0.110	0.004	0.101	0.003
Pupil Diameter (mm)	4.033	0.208	4.308	0.237

Table 8. Means for the target manipulation for the tracking task among the pilots.

Variable	One Target	Two Target
Center Score	144.872	133.197
Score Foot Print	598.031	553.277
Secondary Score	186.250	168.229
Total Score	929.153	854.703
Subjective Workload (TLX)	17.969	40.937

Table 9. Means for the route manipulation for the tracking task.

Variable	Country		City	
	Mean	**SE**	**Mean**	**SE**
Center Score	141.810	1.399	136.152	2.398
Score Foot Print	590.922	5.229	557.333	12.418
Secondary Score	187.719	4.897	178.698	6.575
Total Score	920.913	10.616	872.793	15.892
Subjective Workload (TLX)	38.510	4.085	44.306	4.304
Heart Rate (Beats / Minute)	74.766	3.689	74.397	3.952
Heart Rate Variability (Hz)	0.026	0.122	0.026	0.124
Inspiration Flow	2.284	0.152	2.262	0.160
Inspiration Time (s)	0.230	0.023	0.225	0.026
Respiration Amplitude	519.805	1.264	519.524	1.352
Respiration Cycle Time (s)	3.124	0.095	3.110	0.103
Respiration Duty Cycle Time (s)	0.416	0.004	0.415	0.005
Mean Pitch (Hz)	144.483	14.293	145.592	13.944
Pitch Variance (Hz)	103.687	0.401	103.986	0.381
Maximum Pitch (Hz)	173.805	18.642	173.847	18.101
Pitch Range (Hz)	41.915	0.203	41.542	0.195
Speaking Rate (syllables / second)	4.318	0.146	4.324	0.126
Average Syllable Length (s)	0.160	0.007	0.158	0.006
Average Pause Length (s)	0.135	0.005	0.138	0.006
Percent Pause (%)	37.050	1.378	36.877	1.484
Blink Rate (Blinks / Minute)	7.859	0.264	7.282	0.225
Blink Duration (s)	0.109	0.004	0.102	0.003
Pupil Diameter (mm)	4.155	0.223	4.186	0.220

Table 10. **Means for the route manipulation for the tracking task among the pilots.**

Variable	Country	City
Center Score	141.718	136.352
Score Foot Print	598.559	552.748
Secondary Score	186.563	167.917
Total Score	926.839	857.016
Subjective Workload (TLX)	21.719	37.187

The performance, subjective workload, and physiological data were statistically evaluated using two-way repeated-measures ANOVAs for the tracking task (see Table 11). Note that positively skewed dependent variables (e.g., blink rate) were log transformed prior to analyses. Figures for significant results can be found in Appendix C (tracking), and EEG results are presented in Appendix D.

Table 11. **ANOVA results for the tracking task.**

Variable	**Targets**	**Route**
Center Score	$F(1,11) = 40.24$, $p = 0.0001$*	$F(1,11) = 20.34$, $p = 0.0009$*
Score Foot Print	$F(1,11) = 8.79$, $p = 0.0128$*	$F(1,11) = 9.25$, $p = 0.0112$*
Secondary Score	$F(1,11) = 1.30$, $p = 0.2784$	$F(1,11) = 4.68$, $p = 0.0534$
Total Score	$F(1,11) = 15.45$, $p = 0.0024$*	$F(1,11) = 14.86$, $p = 0.0027$*
Subjective Workload (TLX)	$F(1,11) = 39.74$, $p = 0.0001$*	$F(1,11) = 40.82$, $p = 0.0001$*
Heart Rate (Beats / Minute)	$F(1,11) = 9.99$, $p = 0.0091$*	$F(1,11) = 0.46$, $p = 0.5106$
Heart Rate Variability (Hz)	$F(1,11) = 1.33$, $p = 0.2739$	$F(1,11) = 0.72$, $p = 0.4137$
Inspiration Flow	$F(1,11) = 0.68$, $p = 0.4320$	$F(1,11) = 0.28$, $p = 0.6103$
Inspiration Time (s)	$F(1,11) = 3.81$, $p = 0.0867$	$F(1,11) = 0.48$, $p = 0.5085$
Respiration Amplitude	$F(1,11) = 0.07$, $p = 0.8040$	$F(1,11) = 0.62$, $p = 0.4551$
Respiration Cycle Time (s)	$F(1,11) = 4.33$, $p = 0.0711$	$F(1,11) = 0.34$, $p = 0.5782$
Respiration Duty Cycle Time (s)	$F(1,11) = 1.60$, $p = 0.2420$	$F(1,11) = 0.60$, $p = 0.4592$
Mean Pitch (Hz)	$F(1,11) = 2.98$, $p = 0.1122$	$F(1,11) = 1.51$, $p = 0.2450$
Pitch Variance (Hz)	$F(1,11) = 0.04$, $p = 0.8515$	$F(1,11) = 0.01$, $p = 0.9386$
Maximum Pitch (Hz)	$F(1,11) = 3.98$, $p = 0.0713$	$F(1,11) = 0.00$, $p = 0.9729$
Pitch Range (Hz)	$F(1,11) = 0.00$, $p = 0.9711$	$F(1,11) = 0.29$, $p = 0.6012$
Speaking Rate (syllables / second)	$F(1,11) = 0.50$, $p = 0.4951$	$F(1,11) = 0.01$, $p = 0.9177$
Average Syllable Length (s)	$F(1,11) = 0.64$, $p = 0.4396$	$F(1,11) = 0.26$, $p = 0.6180$
Average Pause Length (s)	$F(1,11) = 0.18$, $p = 0.6776$	$F(1,11) = 1.54$, $p = 0.2406$
Percent Pause (%)	$F(1,11) = 16.32$, $p = 0.0019$*	$F(1,11) = 0.12$, $p = 0.7395$
Blink Rate (Blinks / Minute)	$F(1,11) = 3.52$, $p = 0.0873$	$F(1,11) = 0.46$, $p = 0.5104$
Blink Duration (s)	$F(1,11) = 49.57$, $p = 0.0001$*	$F(1,11) = 17.07$, $p = 0.0017$*
Pupil Diameter (mm)	$F(1,11) = 26.71$, $p = 0.0003$*	$F(1,11) = 2.09$, $p = 0.1765$

Note: * = $p < .05$; Significant interactions are described in the text.

Performance interactions. A significant interaction was present between the route type and number of HVTs for the center score $F(1, 11) = 6.02$, $p < .05$, footprint score $F(1, 11) = 11.46$, $p < .01$, and total performance score $F(1, 11) = 7.59$, $p < .05$. The average center score containing country routes declined from 149.69 in conditions with one HVT to 133.93 in conditions with two HVTs, a difference of 15.76. In contrast, the average center score containing city routes declined from 148.35 in conditions with one HVT to 123.95 in conditions with two HVTs, a difference of 24.40. This differential decline in performance scores was also present in the footprint score. The average footprint score in country routes declined from 597.08 with one HVT to 584.77 with two, a difference of 12.31, while that of city routes declined from 591.71 with one HVT to 522.96 with two, a difference of 68.75. Not surprisingly, the interaction found in the total performance score mirrors the effect seen in the individual scoring components above. The average total score in conditions containing country routes declined from 940.07 with one HVT to 901.76 with two, a difference of 38.31. The average total score in city conditions declined from 920.03 with one HVT to 825.55 with two, a difference of 94.48.

The explanation for these interactions is straightforward. Adding a second HVT in the city increased workload to a greater degree than a second HVT in the country. That is, two target country conditions consisted of two targets in the country, whereas two target city conditions consisted of two targets in the city. Thus, because city targets were more difficult to track than country targets, the effect of doubling the number of targets to track resulted in a differential increase in workload depending on if that target was in the country or city.

Subjective workload interaction. Although not significant, the interaction between route and the number of HVTs trended in the expected direction, $F(1,11) = 4.04$, $p = .07$. The average workload in country routes increased from 33.66 in one HVT conditions to 43.36 in two HVT conditions, a difference of 9.69. The workload in city routes increased from 38.07 in conditions with one HVT to 50.54 in those with two, a difference of 12.47. The interpretation of this interaction is also straightforward, as it occurred for the same reason as the performance interactions. That is, a second HVT in the city increased workload to a greater degree than a second HVT in the country.

Cardiac interaction. There was a significant route by targets interaction for HR, $F(1,11) = 5.61$, $p = .037$. HR in the country conditions increased from 74.4 bpm with one target to 75.1 bpm with two targets, a difference of 0.7 bpm. HR in city conditions increased from 73.0 bpm with one target to 75.8 bpm with two, a difference of 2.9 bpm. This is an interesting interaction, as the data suggests that heart rate was sufficiently sensitive to reflect the particularly high workload conditions of tracking two targets in the city.

Ocular interaction. A significant interaction was also found between the route type and number of HVTs for pupil diameter, $F(1,11) = 7.80$, $p < .05$. In country conditions, dilation increased

from one HVT (4.037) to two (4.273) by 0.236 mm. In contrast, in city conditions, dilation increased from one HVT (4.029) to two (4.342) by 0.313 mm. Thus, the data suggests that pupil dilatation rate was sufficiently sensitive to reflect the particularly high workload conditions of tracking two targets in the city.

Cortical measures. The EEG measures (power at each site and frequency band) were analyzed for each manipulation, but due to the nature of the tracking task, results will not be reported. When tracking two targets, participants had to regularly shift their gaze from one video feed to the other. This caused large saccades to occur at a high rate. Based on examination of time history data, saccades were occurring approximately every 0.9 seconds. So, using four second windows, as was done with the surveillance task, was not possible.

4.3 Tracking Model Results

In order to evaluate the workload models, the relationships between the model outputs (averaged over the course of each trial) and Average TLX are examined (see Table 12). Model 1, Model 2, and Model 4 demonstrated a relationship with subjective workload.

Table 12. Pearson partial correlations between average trial output from ANN models and average subjective workload (TLX) in the tracking task.

Model	R-value
1	.76***
2	.61***
3	.09
4	.36*

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

5.0 DISCUSSION

Researchers and engineers are continually striving to find a solution to meet increasing demand for RPA operations. Future control stations, for instance, are envisioned in which single operators can control multiple aircraft (Dixon et al., 2004). Such systems would allow an efficient use of human resources during low workload operations. A concern, however, is that workload could become excessive due to increased mental demand from managing multiple aircraft, potentially leading to performance decrements and mission failure. One solution to address excessive workload from controlling multiple vehicles, as well as existing challenges RPA operators experience, is to monitor operator state in real-time so that mental overload can be identified and mitigated in a timely fashion. That is, accurate workload assessment would allow the implementation of augmentation strategies *before* performance decrements occur. By examining the feasibly of using physiological measures to monitor workload, and examining the

effectives of several ANN models, this project advances the literature toward real-time workload assessment in field operations.

5.1 Model Performance

There was a stark difference in model results between the surveillance and tracking task. None of the models correlated significantly with subjective workload measures for the surveillance task. These results, however, are not especially surprising in light of the physiological results from the surveillance task, most of which did not reach levels of statistical significance.

Model performance was more promising in the tracking task. Model 1 demonstrated the strongest relationship with subjective workload, followed by Model 2. In a sense, it was surprising that Model 2 performance was inferior to Model 1, given that the only difference between the two was that Model 2 trained on-line. The current instantiations of these on-line training models, however, have room for improvement in that the weights were not saved from session to session. The benefits of online training could take much longer than a few trials to be fully realized. Thus, on-line training should not be viewed in a negative light based on these results, though clearly further research is needed and planned. Models 3 and 4 (which utilize Fz theta only for signal injection instead of the Fz / Pz EEG ratio used in Models 1 and 2) performed poorly by comparison. Model 3 did not significantly correlate with subjective workload, and Model 4 (which trained on-line) did, but to a lesser extent than Models 1 and 2. It is interesting that on-line training improved performance for model 4 but not model 2.

It is not surprising that the models performed better in the tracking task than the surveillance task for several reasons. First, the models were trained using data from a task that was more similar to the target tracking task (see Durkee et al., 2013). Thus, the models would be more likely to detect physiological reflections of workload unique to tracking tasks. In addition, results from the current experiment suggest that several physiological measures, which were included in the model training dataset, were sensitive to the workload manipulations in the tracking task, but not the surveillance task (i.e., pupil diameter and heart rate). Consequently, it would be expected that these vectors would improve model performance in the tracking task, but not necessarily the surveillance task.

Overall, none of the models performed well enough to be implemented into field operations. Models used to drive adaptive automation need to consistently predict operator workload with a high level of resolution. The goal in the current study was to have a model that correlated with subjective workload with an r-value of .85 or greater. Model 1, the best performing model in the current research, demonstrated promising results in the tracking task. The current performance of even this model, however, is still too low for applied use.

The finding that Model 2 performed worse than Model 1 indicates that there is significant room for improvement in the on-line training process. Models 3 and 4 will need to be further evaluated as well. The inferior performance of these models may indicate that using Fz alone for signal injection is inferior to using the Fz / Pz EEG ratio.

Thus, further model improvements are needed and planned before one of the ANN models can be transitioned into an operational setting. Currently, another model development study is underway with the goal of improving model performance. Several new model configurations and inputs will be investigated. In addition, weights from on-line training will be saved for each participant throughout the study. This will increase the volume of participant-specific data available for model training, which should improve model performance. In addition, the models will be task-specific in future research efforts (there will be separate models for the surveillance and tracking tasks).

5.2 Surveillance Task Discussion

In regards to the experimental manipulations for the surveillance task, results indicated that the fuzz manipulation did not significantly impact performance or subjective workload. It was anticipated that the presence of fuzz would make it more difficult for participants to distinguish HVTs from distractors. It could be the case that the fuzz did not sufficiently obscure the visual cues necessary to identify the sniper rifle carried by the HVTs. The distractor manipulation significantly impacted workload and performance.

Physiological measures generally did not demonstrate sensitivity to workload in the surveillance task. Inspiration flow was significantly lower in fuzz conditions than clear conditions, which was in the opposite direction expected based on prior research (Veltman & Gaillard, 1998). There were problems associated with the consistent application of the respiration device.

Pitch variance was greater in high distractor conditions, as would be expected based on prior research (Brenner et al., 1994; Lippold, 1971), but this difference did not reach conventional levels of statistical significance. Overall, the lack of significant physiological findings in the surveillance task was surprising. The evidence from this study is insufficient to advocate the use of inspiration flow or pitch variance in monitoring workload.

5.3 Tracking Task Discussion

For the tracking task, both the route and number of HVTs manipulations significantly impacted workload, performance, and several physiological measures. In regard to the route manipulation, blink duration was significantly reduced in city conditions. Blink duration appears to be a durable measure of workload, as it was identified as an indicator of workload in prior research

using the same tracking task environment (Hoepf et al., 2015), as well as other domains (e.g., Fogarty & Stern, 1989).

In regard to the HVT manipulation, heart rate, pupil diameter, blink duration, and percent pause were sensitive to changes in workload. Further, blink rate, inspiration time, respiration cycle time, and maximum pitch also trended in the direction expected, although the differences did not reach statistical significance. It was a bit surprising that heart rate, but not heart rate variability, was sensitive to changes in workload in this experiment, whereas the opposite was true in a prior experiment using the same tracking task environment (Hoepf et al., 2015). Blink rate, blink duration, and pupil diameter were, however, significant or at least trending in the expected direction in this experiment and the previous experiment, thereby further increasing confidence that these are durable physiological measures for monitoring workload. Changes in percent pause were significant, but the means were in the opposite direction expected. More research is needed to explain these findings.

5.4 Competition Evaluation

An important question in the current research was the extent to which participants remained engaged in the task. To assess participant engagement, a debriefing session with each participant was conducted at the end of the final day of data collection. Every participant indicated that the competition was a motivating factor for them, and that they did not give up on any trials. Thus, it was concluded that the competition was successful and the participants were engaged in the task. In future studies, engagement will be rated on a visual analog scale for each trial in order to quantifiably verify level of engagement.

5.5 Pilots vs. Non-Pilots

An important research question is how non-pilot participants (mostly college students) compare to pilots. After completing data collection from 12 non-pilot participants, data were collected from two pilots. The two pilots did not constitute a sufficient sample size for statistical evaluation. Subjective workload, however, was lower for the pilots than the non-pilots, except in the hardest tracking condition (two targets in the city; see Appendix C). Further investigation on this topic is warranted, as there may be implications if pilots systematically report lower workload than non-pilots. It should be noted that it is possible that the pilots genuinely rated workload lower in the laboratory experiment due to their experience in operational settings.

5.6 Limitations

A limitation in the tracking task is that interpretation of the EEG measures is difficult due to eye artifacts. Alpha power, for instance, increased at several sites in two HVT conditions, which were shown to be the more difficult conditions as evidenced by the performance and subjective workload data. One possible explanation is that EOG artifacts (Fatourechi, Bashashati, Ward, & Birch, 2007) were present in the EEG data due to the additional saccades associated with two target tracking conditions. Due to task demands, participants constantly shifted their gaze between two video feeds in the two target conditions. A very similar finding was present in a previous experiment using the same tracking task environment (Hoepf et al., 2015).

In a sense, it is encouraging the EEG results are consistent between two studies that used the same task. In fact, it is likely that the neural network architecture of the workload models capitalized on these eye artifacts within the EEG data to identify saccadic eye movement patterns consistent with increased tracking activity. This of course is concerning for two reasons. First, if the workload models are in fact using eye artifacts in EEG, this would likely dilute the performance of the models in other task environments. Indeed, this may have been partially responsible for the poor model performance in the surveillance task.

Attempts were made to compute artifact-free data by reducing the window size. A small window size (0.53 seconds) was tested in attempt to analyze "clean" segments of data between saccades (see Appendix D). However, the associated spectral resolution of the EEG data is very low, and thus unreliable. The usefulness of EEG data collected and new methods for removing frequent ocular artifacts is being considered.

Another limitation of the current study is the small pilot sample size. It would have been preferable to collect data from twelve or more pilots in order to statistically compare the two samples. This would add confidence to conclusions regarding potential differences between pilots and non-pilots. In general, however, the addition of the pilots strengthened the study by providing a glimpse as to how pilots may compare to non-pilots. In addition, debriefing discussions with the pilots also revealed that the task environment used in this study was realistic, which was also important.

5.7 Implications and Future Research

The physiological measures used in the current investigation generally showed more sensitivity to changes in workload in the tracking task than the surveillance task. This finding indicates that physiological measures that are well suited for one task environment are not necessarily suited for other task environments. Thus, researchers should not assume that the physiological measures that demonstrated sensitivity to workload in the tracking task of this investigation will reflect

workload in other task environments. Before physiological measures can be utilized in field operations, researchers will need to conduct studies evaluating the effectiveness of the specific physiological measures to be used in those field operations using artificial task environments that accurately reflect the nature of the task.

Researchers (e.g., Wilson, 1992) have suggested that the stress of real-world operations could result in increased physiological responses compared to laboratory experiments. Heart rate, for instance, did not reflect changes due to the workload manipulations in the surveillance task of the current research. However, heart rate could demonstrate sensitivity to workload in RPA surveillance field operations. The findings of the current investigation, though based on a realistic task environment, should not be expected to map perfectly onto field operations because the lab setting may not reflect the stress of field operations. Laboratory research with improved modeling capabilities need to be conducted before transitioning to the field.

An important goal of this line of research is to make real-time assessments of operator workload for the purpose of augmenting performance. In the future, researchers should explore physiologically-based adaptive automation, which is a method of providing assistance to operators by introducing automation only when it is required (Parasuraman, Mouloua, & Molloy, 1996; Scerbo, 1996). Wilson and Russell (2007), for example, used physiological features to train an artificial neural network to classify workload, which in turn was used to determine when the operator needed assistance. The researchers demonstrated a performance improvement of approximately 50% by using the adaptive automation technique. Future research will focus on improving the effectiveness of workload assessment models in RPA task environments using physiological measures. New physiological features will be added to the array of model inputs and improvements to the existing physiological features will be made. An enhanced modeling approach using advanced software will be implemented. Lastly, model training will be improved to produce models that are individualized to the person and the task.

6.0 CONCLUSIONS

With the increasing use of RPA in military operations, research is needed to address the performance of our Airman in these operational domains. The current study implemented an operationally realistic RPA surveillance and tracking task, which was used to investigate performance, subjective workload, and physiological data. Utilizing the SAA taxonomy as a framework of research, the current study focuses on sensing the operator's cognitive state using physiological measures, and then attempts to assess that state using ANN models.

Several physiological features were identified that show potential for monitoring workload in RPA operations in real-time. Overall, these results are encouraging in that the physiological

measures clearly demonstrated sensitivety to workload in the tracking task. Thus, the current investigation takes a step toward physiologically-based workload assessment in RPA operations.

The primary goal of this research was to evaluate four ANN models. Several of the ANN models significantly correlated with subjective workload in the tracking task, but not the surveillance task. Unfortunately, the correlations were not strong enough to drive adaptive automation. Further model development is needed before focusing on the augment phase of the SAA taxonomy. Furthermore, improvements to modeling techniques and testing are required before field implementation will be possible. Such model improvements are underway.

7.0 REFERENCES

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*, 276-292.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews, 44,* 58-75. doi: 10.1016/j.neubiorev.2012.10.003

Brenner, M, Doherty, E. T., & Shipp, T. (1994). Speech measures indicating workload demand. *Aviation, Space and Environmental Medicine, 65*, 21-26.

Christ, R. E., Hill., S. G., Ayers, J. C., Iavecchia, H. M., Zaklad, A. L., & Bittner, A. (1993). *Application and validation of workload assessment techniques* (Technical Report 974). Alexandria, VA: U.S. Army Research Institute for the Behavioral Sciences.

Christensen, J. C., Estepp, J. R., Wilson, G. F., & Russell, C. A. (2012). The effects of day-to-day variability of physiological data on operator functional state classification. *NeuroImage, 59,* 57-63. doi: 10.1016/j.neuroimage.2011.07.091

Cohen, H. D., Goodenough, D. R., Witkin, H. A., Oltman, P., Gould, H., & Shulman, E. (1975). The effects of stress on components of the respiration cycle. *Psychophysiology, 12,* 377-380. doi: 10.111/j.1469-8986.1975.tb00005.x

Collier, R. (1974). Laryngeal muscle activity, subglottal air pressure, and the control of pitch in speech (*Status Report on Speech Research SR-39/40).* New Haven, Conn: Haskins Laboratories.

Credlebaugh, C., Middendorf, M., Hoepf, M., & Galster, S. (2015). EEG data analysis using artifact separation. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology* (pp. 434-439). Dayton, OH: Wright State University.

Dixon, S. R., Wickens, C. D., & Chang, D. (2004). Unmanned aerial vehicle flight control: False alarms versus misses. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, New Orleans, LA, 48,* 152-156. doi: 10.1177/154193120404800133

Durkee, K., Geyer, A., Pappada, S., Ortiz, A., & Galster, S. (2013). Real-time workload assessment as a foundation for human performance augmentation. *Proceedings of the 7th HCI International Conference, NV, 8027*, 279-288. doi: 10.1007/978-3-642-39454-6_29

Durkee, K., Pappada, S., Ortiz, A., Feeney, J., & Galster, S. (2015). Using context to optimize a functional state estimation engine in unmanned aircraft system operations. In *Foundations of Augmented Cognition* (pp. 24-35). Springer International Publishing. doi: 10.1007/978-3-319-20816-9_3

Epling, S., Middendorf, M., Hoepf, M., Gruenwald, C., Stork, L., & Galster, S. (2015). The electrooculogram and a new blink detection algorithm. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology* (pp. 512-517). Dayton, OH: Wright State University.

Fatourechi, M., Bashashati, A., Ward, R. K., & Birch, G. E. (2007). EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology, 118,* 480-494. doi: 10.1016/j.clinph.2006.10.019

Fiorita, A. L., Middendorf, M. S., & McMillan, G. R. (1992). Maintaining subject motivation in long-term experiments using performance incentives and penalties. *Proceedings of the 36th Annual Meeting of the Human Factors Society,* Volume 2, 1335-1339. doi: 10.1518/107118192786749423

Fogarty, C. & Stern, J. (1989). Eye movements and blinks: Their relationship to higher cognitive processes. *International Journal of Psychophysiology, 8,* 35-42. doi: 10.1016/0167-8760(89)90017-2

Galster, S. M., & Johnson, E. M. (2013). Sense-assess-augment: A taxonomy for human effectiveness (Report No. AFRL-RH-WP-TM-2013-0002). Wright-Patterson Air Force Base: Air Force Research Laboratory, Human Effectiveness Directorate.

Gevins, A., Smith, M. E., Leong, I. I., McEvoy, L., Whitfield, S., Du, R., & Rush, G. (1998). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Human Factors, 40,* 79-91. doi: 10.1518/001872098779480578

Grossman, P. (1992). Respiratory and cardiac rhythms as windows to central and autonomic biobehavioral regulation: Selection of window frames, keeping the panes clean and viewing the neural topography. *Biological Psychology, 34,* 131-161. doi: 10.1016/0301-0511(92)90013-k

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine, 69,* 360-367.

Hart, S. G., & Staveland, L. E. (1988). Development of the NASA- TLX (Task Load Index): Results of experimental and theoretical research. *Advances in Psychology, 52,* 139-183. doi: 10.1016/S0166-4115(08)62386-9

Hendy, K. C., Hamilton, K. M., & Landry, L. N. (1993). Measuring subjective workload: When is one scale better than many? *Human Factors: The Journal of the Human Factors and Ergonomics Society, 35,* 579-601. doi: 10.1177/001872089303500401

Hoepf, M., Middendorf, M., Epling, S. & Galster, S. (2015). Physiological indicators of workload in a remotely piloted aircraft simulation. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology* (pp. 428-433). Dayton, OH: Wright State University.

Jasper, H. (1958). Report of the committee on methods of clinical examination. *Electroencephalography and Clinical Neurophysiology, 10*, 370-375. doi: 10.1016/0013-4694(58)90053-1

Jorna, P. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology, 34*, 237-257. doi: 10.1016/0301-0511(92)90017-O

Lippold, O. (1971). Physiological tremor. *Scientific American, 224*, 65-73.

Middendorf, M., Gruenwald, C., Stork, L., Epling, S., Hoepf, M., & Galster, S. (2015). Saccade detection using polar coordinates – A new algorithm. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology* (pp. 518-523). Dayton, OH: Wright State University.

Mulder, L. J. M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology, 34*, 205-236. doi:10.1016/0301-0511(92)90016-N

Niedermeyer, E. & Da Silva, F. L. (2004). *Electroencephalography: Basic principles, clinical applications, and related fields*. Lippincot Williams & Wilkins.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors, 33*, 17-31. doi: 10.1177/001872089103300102

Onton, J., Delorme, A., & Makeig, S. (2005). Frontal midline EEG dynamics during working memory. *NeuroImage, 27*, 341-356. doi:10.1016/j.neuroimage.2005.04.014

Parasuraman, R., Mouloua, M. & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 38*, 665-679. doi: 10.1518/001872096778827279

Pfurtscheller, G. & Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology. 110*, 1842-1857. doi: 10.1016/S1388-2457(99)00141-8

Picton, T., Roon, P., Armilio, M., Berg, P., Ille, N., & Scherg, M.(2000). The correction of ocular artifacts: A topographical perspective. *Clinical Neurophysiology, 111*, 53-65. doi: 10.1016/S1388-2457(99)00227-8

Pomplun, M. & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. In D. Harris, V. Duffy, M. Smith, & C. Stephanidis (eds.), *Human-Centred Computing: Cognitive, Social, and Ergonomic Aspects.* Vol. 3 of the *Proceedings of the 10th International Conference on Human-Computer Interaction, HCI 2003*, Crete, Greece.

Porges, S. W., & Byrne, E. A. (1992). Research methods for measurement of heart rate and respiration. *Biological Psychology, 34*, 93-130. doi:10.1016/0301-0511(92)90012-J

Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology, 34*(2-3), 259-287. doi:10.1016/0301-0511(92)90018-P

Rose, M. R., Arnold, R. D., & Howse, W. R. (2013). Unmanned aircraft systems selection practices: Current research and future directions. *Military Psychology, 25*, 413-427.

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* [CD-ROM]. Mahwah, NJ: Erlbaum.

Suess, W. M., Alexander, A. B., Smith, D. D., Sweeney, H. W., & Marion, R. J. (1980). The effects of psychological stress on respiration: A preliminary study of anxiety and hyperventilation. *Psychophysiology, 17*, 535-540. doi: 10.1111/j.1469-8986.1980.tb02293.x

U.S. Department of Defense. (2011). *Unmanned systems integrated roadmap FY2011-2036* (Reference No. 11-S-3613). Washington, DC: Department of Defense.

Veltman, J. A., & Gaillard, A. W. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics, 41*, 656-669. doi: 10.1080/001401398186829

Wilson, G. F. (1992). Applied use of cardiac and respiration measures: Practical considerations and precautions. *Biological Psychology, 34*, 163-178. doi:10.1016/0301-0511(92)90014-L

Wilson, G. F. (2001). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *International Journal of Aviation Psychology, 12*, 3-18. doi: 10.1207/S15327108IJAP1201_2

Wilson, G. F., & Russell, C. A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 49*, 1005-1018. doi: 10.1518/001872007X249875

Young, M. S., & Stanton, N. A. (2002). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science, 3*, 178-194. doi: 10.1080/14639220210123789

Zander, T. O., & Kothe, C. (2011). Towards passive brain–computer interfaces: applying brain-computer interface technology to human–machine systems in general. *Journal of Neural Engineering, 8*(2), 1-5. doi: 10.1088/1741-2560/8/2/025005

Zarjam, P., Epps, J., & Lovell, N. H. (2012). Characterizing mental load in an arithmetic task use entropy-based features. In *The 11th International Conference on Information Sciences, Signal Processing, and their Applications: Main Tracks (pp.199-204).* doi: 0.1109/ISSPA.2012.6310545

APPENDIX A – SCREENSHOTS

Surveillance (Low Distractors, No Fuzz)

Surveillance (High Distractors, No Fuzz)

Surveillance (Low Distractors, Fuzz)

Surveillance (High Distractors, Fuzz)

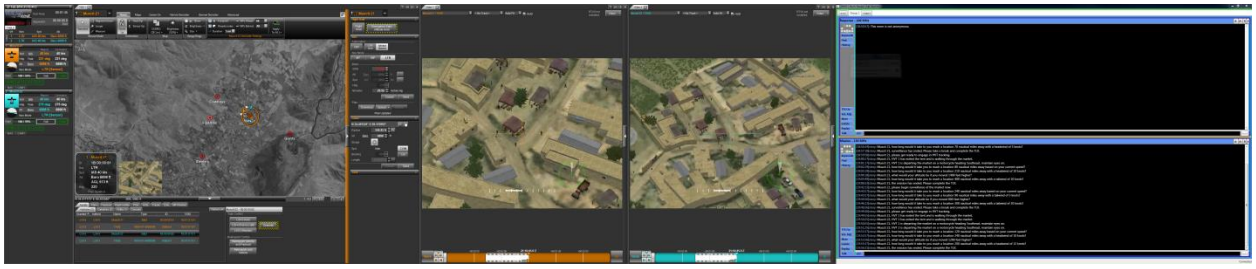
Tracking (One Target, Country)

Tracking (One Target, City)

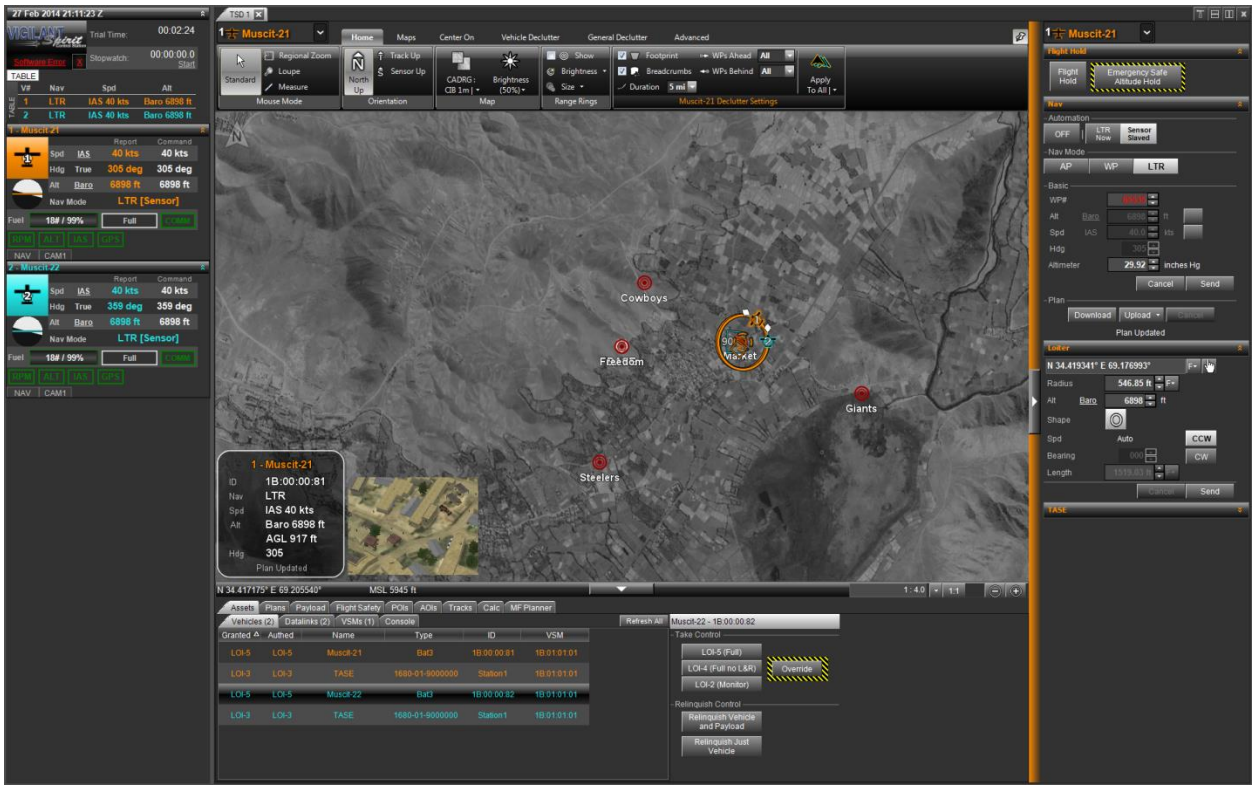
Tracking (Two Targets, Country)

Tracking (Two Targets, City)

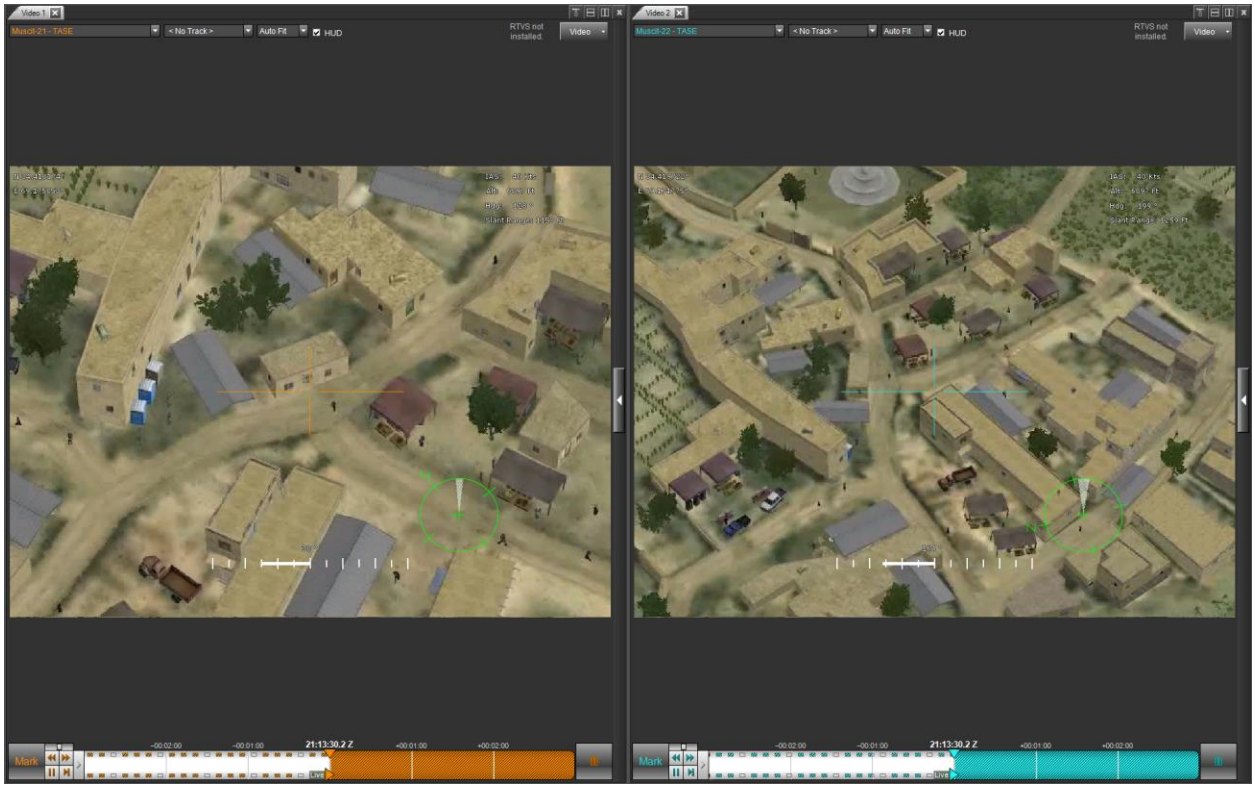
Overall Display (Includes Vigilant Spirit on left and middle monitors and MMC on the right monitor)



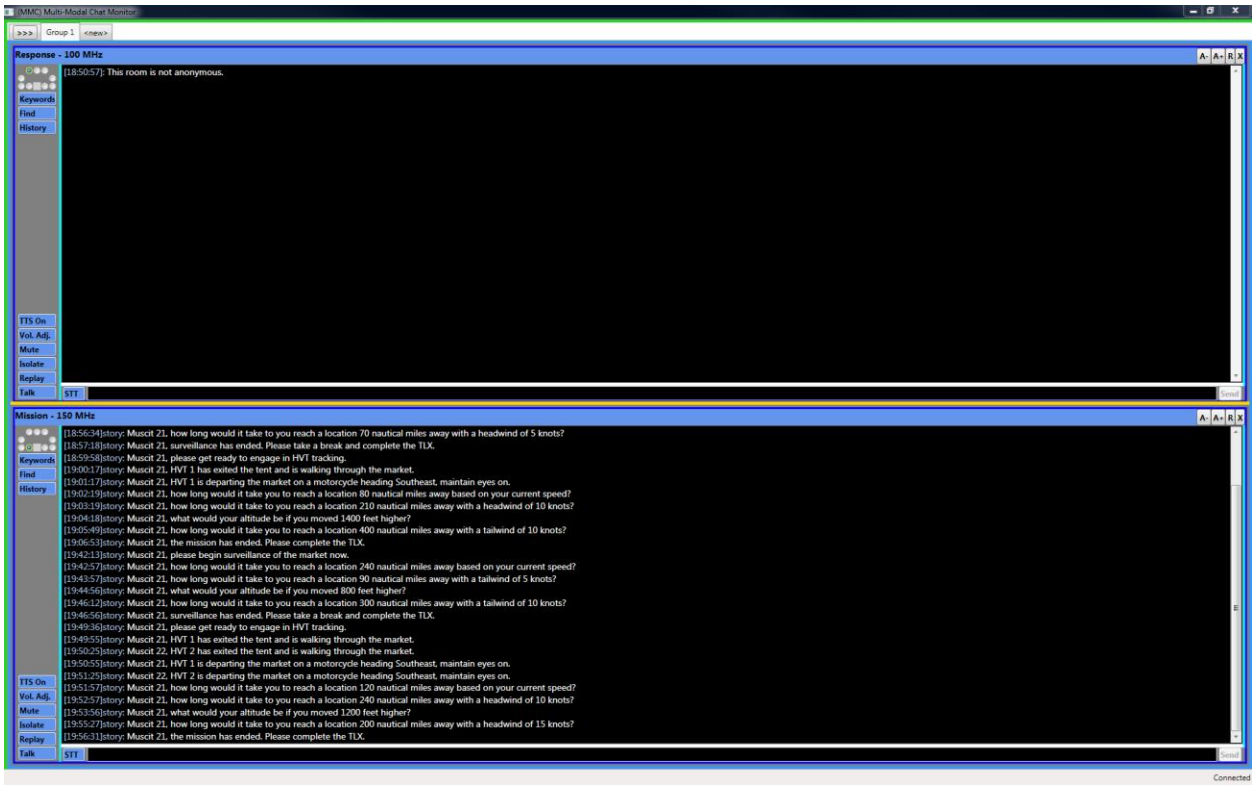
Tactical Situation Display



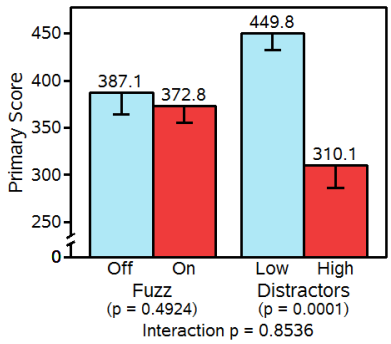
Sensors

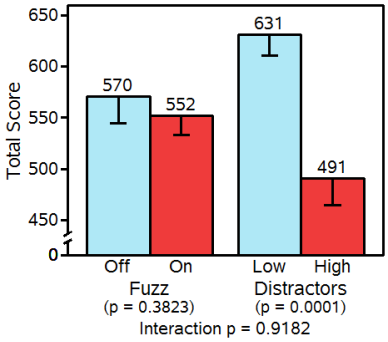


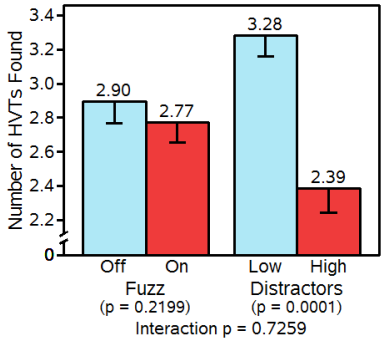
Multi-Modal Communication Tool

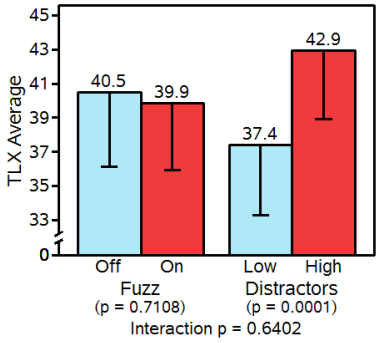


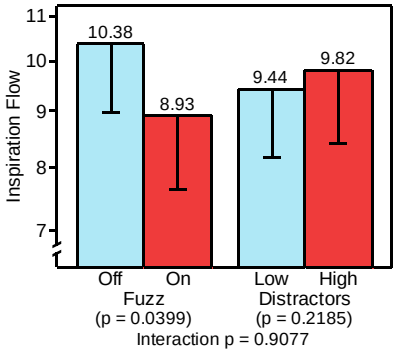
APPENDIX B – SIGNIFICANT RESULTS (SURVEILLANCE)



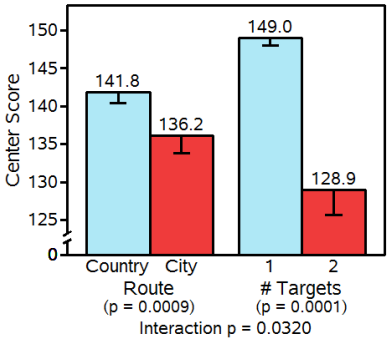


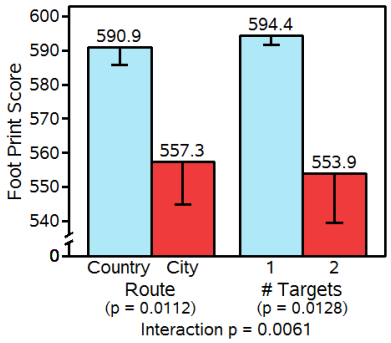


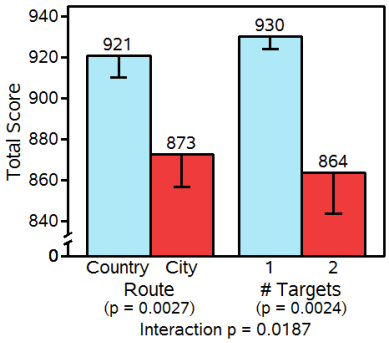


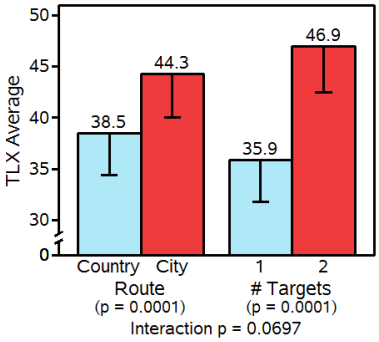


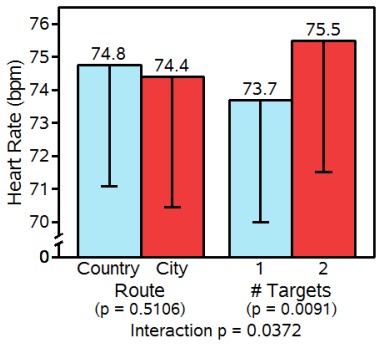
APPENDIX C – SIGNIFICANT RESULTS (TRACKING)

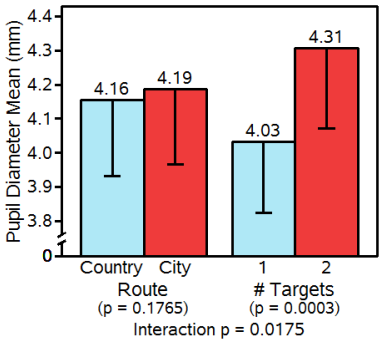


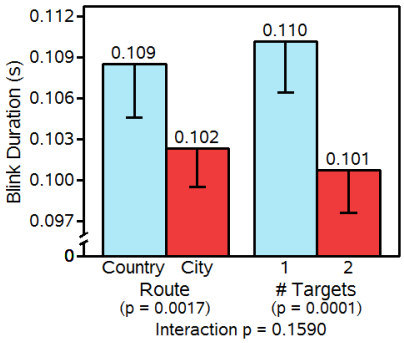


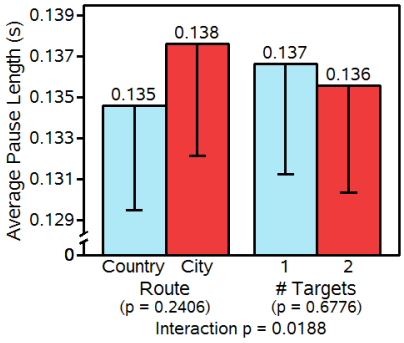


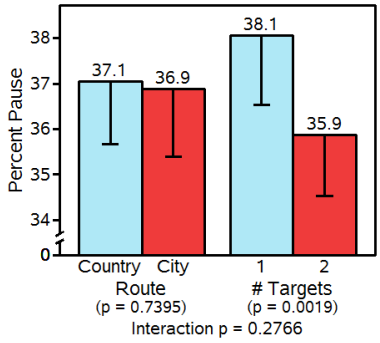


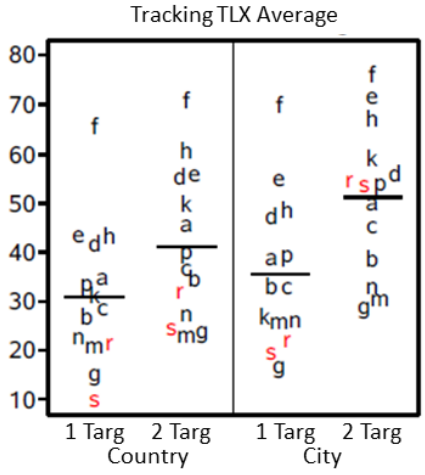






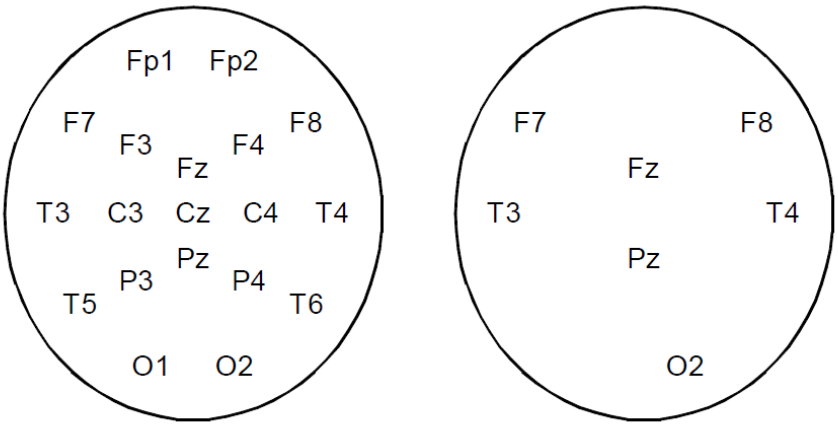






TLX Average Scatter Plot-Tracking. Average TLX scores are identified above for the tracking task. Each letter represents a participant. Letters "r" and "s" are the two pilots that participated in the study. Note that they rated subjective workload lower than most participants, except for in the hardest condition (2 targets, city).

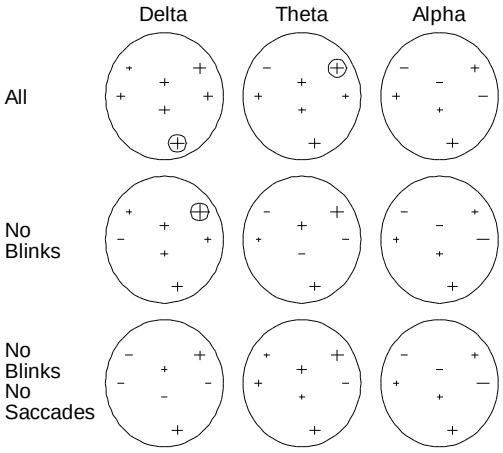
APPENDIX D – EEG REFERENCE



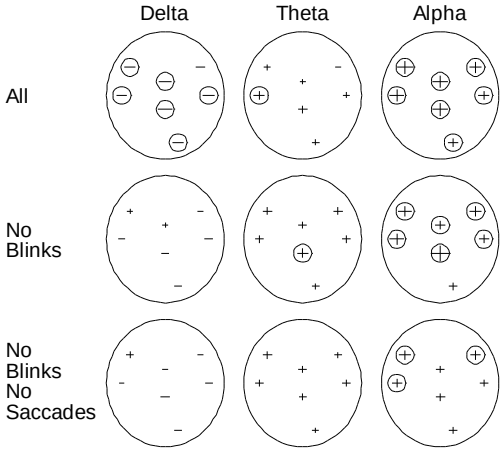
All sites (left) and sites used in this study (right)

EEG RESULTS - SURVEILLANCE

Fuzz: On - Off



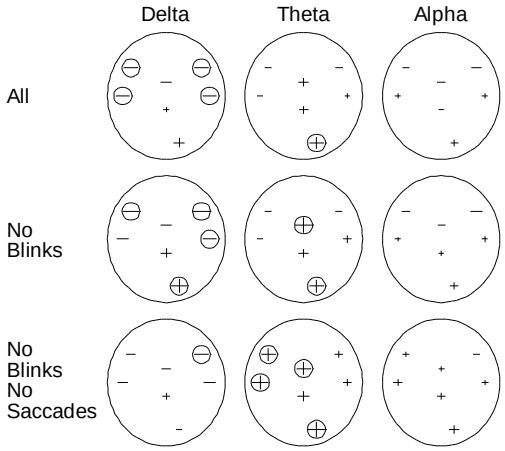
Distractors: High - Low



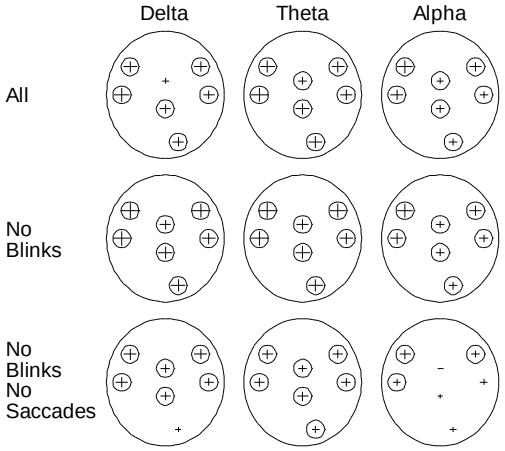
Main effect tests for Surveillance. The sign is the direction of the difference in log power (i.e., a plus sign means more power for on than for off and more power for high than for low). The size of the sign is relative absolute value of the *t* statistic (i.e., larger sign means smaller *p*-value). If the sign is circled then $p \leq 0.05$.

EEG RESULTS - TRACKING

Route: City - Country



Targets: 2 - 1



Main effect tests for Tracking. The sign is the direction of the difference in log power (i.e., a plus sign means more power for city than for country and more power for 2 targets than for 1). The size of the sign is relative absolute value of the *t* statistic (i.e., larger sign means smaller *p*-value). If the sign is circled then $p \leq 0.05$.

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	artificial neural network
ANOVA	analysis of variance
ANS	autonomic nervous system
dB	decibel
ECG	electrocardiography
EEG	electroencephalography
EOG	electrooculography
HEOG	horizontal electrooculography
HR	heart rate
HRV	heart rate variability
HUD	heads-up display
HVT	high value target
Hz	Hertz
MMC	Multi-Modal Communication tool
RPA	remotely piloted aircraft
SAA	Sense-Assess-Augment
SE	standard error
SME	subject matter expert
TLX	Task Load Index
VEOG	vertical electrooculography
VS	Vigilant Spirit